Clustering is difficult only when it does not matter*

Amit Daniely † Nati Linial ‡ Michael Saks §

May 23, 2012

Abstract

Numerous papers ask how difficult it is to cluster data. We suggest that the more relevant and interesting question is how difficult it is to cluster data sets *that can be clustered well*. More generally, despite the ubiquity and the great importance of clustering, we still do not have a satisfactory mathematical theory of clustering. In order to properly understand clustering, it is clearly necessary to develop a solid theoretical basis for the area. For example, from the perspective of computational complexity theory the clustering problem seems very hard. Numerous papers introduce various criteria and numerical measures to quantify the quality of a given clustering. The resulting conclusions are pessimistic, since it is computationally difficult to find an optimal clustering of a given data set, if we go by any of these popular criteria. In contrast, the practitioners' perspective is much more optimistic. Our explanation for this disparity of opinions is that complexity theory concentrates on the worst case, whereas in reality we only care for data sets that can be clustered well.

We introduce a theoretical framework of clustering in metric spaces that revolves around a notion of "good clustering". We show that if a good clustering exists, then in many cases it can be efficiently found. Our conclusion is that contrary to popular belief, clustering should not be considered a hard task.

Keywords: Cluster Analysis, Hardness of clustering, Theoretical Framework for clustering, Stability.

*Credit for this title goes to Tali Tishby who stated this in a conversation with one of us many years ago.

†Department of Mathematics, Hebrew University, Jerusalem 91904, Israel. Supported in part by a binational Israel-USA grant 2008368. amit.daniely@math.huji.ac.il

‡School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel. Supported in part by a binational Israel-USA grant 2008368. nati@cs.huji.ac.il

§Department of Mathematics, Rutgers University, Piscataway, NJ 08854. Supported in part by NSF under grant CCF-0832787 and by a binational Israel-USA grant 2008368. saks@math.rutgers.edu.

1 Introduction

Clustering is the task of partitioning a set of objects in a meaningful way. Notwithstanding several recent attempts to develop a theory of clustering (e.g. [1, 4, 9]), our foundational understanding of the matter is still quite unsatisfactory.

The clustering problem deals with a set of objects X that is equipped with some additional structure, such as a dissimilarity (or similarity) function $w : X \times X \to [0, \infty)$. Informally, we are seeking a partition of X into clusters, such that objects are placed in the same cluster iff they are sufficiently similar. Here are some concrete popular manifestations of this general problem.

1. A very popular optimization criterion is k-means. Aside from X and w one is given an integer k. The goal is partition X into k parts $C_1, \ldots, C_k$ and find a *center* $x_i \in C_i$ in each part so as to minimize $\sum_i \sum_{y \in C_i} w^2(y, x_i)$. Other popular criteria of similar nature are k-medians, min-sum and others.

2. Many clustering algorithms work "bottom up". Initially, every singleton in X is considered as a separate cluster, and the algorithm proceeds by repeatedly merging nearby clusters. Other popular algorithms work "top down": Here we start with a single cluster that consists of the whole space. Subsequently, existing clusters get split to improve some objective function.

3. Several successful methods use spectral methods. One associates a matrix (e.g. a Laplacian) to (X, w), and partitions X according to the eigenvectors of this matrix.

Approaches to the clustering problem that focus on some objective function, usually result in NP-hard optimization problems. Consequently, most existing theoretical studies concentrate on designing approximation algorithms for such optimization problems and proving appropriate hardness results.

However, the practical purpose of clustering is *not* to optimize such objectives. Rather, our goal is to find a meaningful partition of the data (provided, of course, that such a partition exists). The point that we advocate is that a satisfactory theory of clustering, should start with a definition of a good clustering and proceed to determine when a good clustering can be found efficiently. In this paper, we follow this approach when the underlying space in a metric[1] space.

This perspective leads to conclusions which are at odds with common beliefs regarding clustering. This applies, in particular, to the computational hardness of clustering. The infeasibility of optimizing most of the popular objectives led many theoreticians, to the bleak view that clustering is hard. However, we show that in many circumstances a good clustering can be efficiently found, leading to the opposite conclusion. From the practitioner's viewpoint, *"clustering is either easy or pointless"* – that is, whenever

[1]The assumption that d is a metric is not too strict. E.g., much of what we do applies even if we weaken the triangle inequality to $\lambda \cdot d(x, z) \leq d(x, y) + d(y, z)$ for λ bounded away from zero.

the input admits a good clustering, finding it is feasible. Our analysis provides some support to this view.

This work is one of several recent attempts to develop a mathematical theory of clustering. For more on the relevant literature, see Section 4.

1.1 A Theoretical Framework for Clustering in Metric Spaces

There are numerous notions of clusters in data sets and clustering methods to be found in the literature. Although not necessarily stated explicitly, these methods are guided by an ideal (in the Platonic sense) notion of a *good cluster* in a space X. This is a subset $C \subseteq X$ such that if $x \in C$ and $y \notin C$, then x is *substantially* closer to C than y is. To rule out trivialities we usually require C to be *big* enough. This, in particular, eliminates the possibility of trivial singleton clusters. Even more emphasis is put on problems of *clustering*. Here we seek partitions of the space X into *clusters* such that every $x \in X$ is *substantially* closer to the cluster containing it than to any other cluster. The problem is specified in terms of a proximity measure $\Delta(x, A)$ between elements $x \in X$ and subsets $A \subseteq X$. Numerous natural choices for $\Delta(\cdot, \cdot)$ suggest themselves. For example, if X is a metric space, it is reasonable to define $\Delta(x, A)$ in terms of x's distances from members of A.

In the present paper we consider a metric space (X, d) from which data points are sampled[2] according to a probability distribution P. The definition we adopt here is $\Delta(x, A) = E_{y \sim P}[d(x, y) | y \in A]$. Other interesting definitions suggest themselves, e.g., $\Delta'(x, A) = \inf_{y \in A \setminus \{x\}} d(x, y)$.

A technical comment: The definition of $\Delta(x, A)$ depends on the distribution P. To simplify notations we omit subscripts such as P when they are clear from the context.

Formally, we say that $C \subset X$ is an (α, γ)**-cluster** for $\alpha > 0$, $\gamma > 1$ if $P(C) \geq \alpha$ and for (almost-)every[3] $x \in C, y \notin C$,

$$\Delta(y, C) \geq \gamma \cdot \Delta(x, C).$$

Likewise, a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ of X is an (α, γ)**-clustering** for some $\alpha > 0$, $\gamma > 1$ if

$$\Delta(x, C_j) \geq \gamma \cdot \Delta(x, C_i)$$

for every $i \neq j$ and (almost-)every $x \in C_i$ and, in addition, $P(C_i) \geq \alpha$ for every i.

A few technical points are in place.

- We study (α, γ)-clusterings of a space as well as partitions of a space into (α, γ)-clusters. We note that although these two notions are similar, they are **not** identical.

[2] In certain cases it is inappropriate to assume that points of X are drawn at random. It is also possible that we do not know how X is sampled. In such circumstances, we consider P as the uniform distribution on X.

[3] *Almost* means, as usual, that we are allowing an exceptional set of measure zero.

- Our results hold if we choose instead to define $\Delta(x, A)$ as $E[d(x, y)|y \in A \setminus \{x\}]$. This definition is perfectly reasonable, but it leads to certain minor technical complications that the current definition avoids. Moreover, the difference between the two definitions is rather insignificant, since our main interest is in cases where $P(\{x\}) \ll P(A)$.

Our main focus here is on efficient algorithms for finding (α, γ)-clusters and clusterings. The analysis of these algorithms rely on the structural properties of such clusters. We can now present our main results. To simplify matters without compromising the big picture, we state our theorems in the case when X is a given finite metric space.

Theorem 1.1 *For every fixed $\gamma > 1, \alpha > 0$ there is an algorithm that finds all (α, γ)-clusterings of a given finite metric space X and runs in time poly$(|X|)$.*

Theorem 1.2 *There is a polynomial time algorithm that on input a finite metric space X and $\alpha > 0$ finds all γ-clusters in X with $\gamma > 3$ and a partition of X into (α, γ)-clusters with $\gamma > 3$, provided one exists. Moreover, the latter problem is NP-hard for $\gamma = 5/2$.*

1.2 An overview

Our discussion splits according to the value of the parameter α. When α is bounded away from zero we work by exhaustive sampling (e.g. as in [2]). We first sample a small set of points S from the space. Since $|S|$ is small (logarithmic in an error parameter), it is computationally feasible to consider all possible partitions Π of S. To each partition Π of S we associate a clustering that can be viewed as the corresponding "Voronoi diagram". If the space has an (α, γ)-clustering $\mathcal{C}$, let Π^* be the partition of S that is consistent with $\mathcal{C}$. We show that the "Voronoi diagram" of Π^* nearly coincides with $\mathcal{C}$ provided that γ is bounded away from 1. Concretely, Lemma 2.2 controls the distances between points that reside in distinct clusters in an (α, γ)-clustering. Together with Hoeffding's inequality this yields Lemma 2.3 and Corollary 2.4 which show that the "Voronoi diagram" of an appropriate partition of a small sample is nearly an (α, γ)-clustering. Lemma 2.5 speaks about the collection of all possible (α, γ)-clusterings of the space. It shows that every two distinct (α, γ)-clusterings must differ substantially. Consequently (Corollary 2.6) there is a bound on the number of (α, γ)-clusterings that any space can have. All of this is then used to derive an efficient algorithm that can find all (α, γ)-clusterings of the space, proving Theorem 1.1.

In section 3 we deal with the case of small α. This affects the analysis, since we require that the dependency of the algorithm's runtime on α be poly$(\frac{1}{\alpha})$. We show that $(\alpha, 3+\epsilon)$-clusters are very simple: Such a cluster is a ball and any two such clusters that intersect are (inclusion) comparable. These structural properties are used to derive an efficient algorithm that partitions the space into $(\alpha, 3+\epsilon)$-clusters (provided that such a partition exists), proving the positive part of Theorem 1.2. To match this result, we

show that finding a partition of the space into $(\alpha, 2.5)$-clusters is NP-Hard, proving Theorem 1.2 in full.

Lastly, in section 4 we discuss some connection to other work, both old and new, as well as some open questions arising from our work.

2 Clustering into Few Clusters – α is bounded away from zero

Throughout the section, X is a metric space endowed with a probability measure P. To avoid confusion, other probability measures that are used throughout, are denoted by Pr. We define a metric d between two collections of subsets of X, say $\mathcal{C} = \{C_1, \ldots, C_k\}$ and $\mathcal{C}' = \{C'_1, \ldots, C'_k\}$. Namely, $d(\mathcal{C}, \mathcal{C}') = \min P(\cup_{i=1}^k C_i \oplus C'_{\sigma(i)})$ where $A \oplus B$ denotes symmetric difference, and the minimum is over all permutations $\sigma \in S_k$. The definition of $d(\mathcal{C}, \mathcal{C}')$ extends naturally to the case where $\mathcal{C}$ and $\mathcal{C}'$ have k resp. l sets and, say $l \leq k$. The only change is that now $\sigma : [l] \to [k]$ is $1:1$.

We define Δ also on sets. If $A, B \subseteq X$, we define $\Delta(A, B)$ as the expectation of $d(x, y)$ where x and y are drawn from the distribution P restricted to A and B respectively. It is easily verified that Δ is symmetric and satisfies the triangle inequality. It is usually *not* a metric, since $\Delta(A, A)$ is usually positive.

Proposition 2.1 *For every $A, B, C \subset X$,*

$$\Delta(A, B) = \Delta(B, A) \quad \textit{and} \quad \Delta(A, B) \leq \Delta(A, C) + \Delta(C, B)$$

As the following lemma shows, distances in an (α, γ)-clustering are fairly regular

Lemma 2.2 *Let $C_1, \ldots, C_k$ be an (α, γ)-clustering and let $i \neq j$. Then*

1. *For almost every $x \in C_i, y \in C_j$, $\frac{\gamma-1}{\gamma}\Delta(y, C_i) \leq d(x, y) \leq \frac{\gamma^2+1}{\gamma(\gamma-1)}\Delta(y, C_i)$*
2. *For almost every $x, y \in C_i$, $d(x, y) \leq \frac{2}{\gamma-1} \cdot \Delta(x, C_j)$*

Proof. Let $x \in C_i, y \in C_j$. For the left inequality in part 1, note that

$$\begin{aligned} d(x, y) &\geq \Delta(y, C_i) - \Delta(x, C_i) \\ &\geq \Delta(y, C_i) - \frac{1}{\gamma} \cdot \Delta(x, C_j) \\ &\geq \Delta(y, C_i) - \frac{1}{\gamma} \cdot [d(x, y) + \Delta(y, C_j)] \\ &\geq \Delta(y, C_i) - \frac{1}{\gamma} \cdot [d(x, y) + \frac{1}{\gamma} \cdot \Delta(y, C_i)] \end{aligned}$$

For the right inequality,

$$\begin{aligned} d(x,y) &\leq \Delta(x,C_i) + \Delta(y,C_i) \\ &\leq \frac{1}{\gamma} \cdot \Delta(x,C_j) + \Delta(y,C_i) \\ &\leq \frac{1}{\gamma} \cdot (d(x,y) + \Delta(y,C_j)) + \Delta(y,C_i) \\ &\leq \frac{1}{\gamma} \cdot (d(x,y) + \frac{1}{\gamma} \cdot \Delta(y,C_i)) + \Delta(y,C_i) \end{aligned}$$

For part 2,

$$\begin{aligned} d(x,y) &\leq \Delta(x,C_i) + \Delta(y,C_i) \\ &\leq \frac{1}{\gamma} \cdot [\Delta(x,C_j) + \Delta(y,C_j)] \\ &\leq \frac{1}{\gamma} \cdot [2 \cdot \Delta(x,C_j) + d(x,y)] \end{aligned}$$

□

Note that for $\gamma \to \infty$ all distances $d(x,y)$ with $x \in C_i$ and $y \in C_j$ are roughly equal and $d(x_1,x_2) \ll d(x_1,y)$ for all $x_1, x_2 \in C_i$ and $y \in C_j$ with $i \neq j$.

We show next how to recover an (α,γ)-clustering by sampling. For $x \in X$ and $A \subseteq X$ finite, we denote the average distance from x to A's elements by $\Delta_U(x,A) := \frac{1}{|A|}\sum_{y\in A} d(x,y)$. A finite sample set S provides us with an estimate for the distance of a point x from a (not necessarily finite) $C \subseteq X$. Namely, we define the **empirical proximity** of x to C as $\Delta_{emp}(x,C) := \Delta_U(x, C \cap S)$.

We turn to explain how we recover an unknown (α,γ)-clustering of X with $\alpha > 0$ and $\gamma > 1$. Consider a collection $\mathcal{C} = \{\mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{k}\} \subseteq \mathrm{X}$ of disjoint subsets of X. We define a "Voronoi diagram" corresponding to S, denoted $\mathcal{C}^\gamma = \{C_1^\gamma, \ldots, C_k^\gamma\}$. Here

$$C_i^\gamma = \{x \in X : \forall j \neq i,\ \gamma \cdot \Delta_{emp}(x,C_i) < \Delta_{emp}(x,C_j)\}.$$

If $\mathcal{C}$ is a (α,γ)-clustering of X, we expect $\mathcal{C}^\gamma$ to be a good approximation of $\mathcal{C}$.

Lemma 2.3 *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be an (α,γ)-clustering of X. Let $S = \{Z_1, \ldots, Z_m\}$ be an i.i.d. sample with distribution P and let $q \neq p$. Then, for every $x \in C_q, \epsilon > 0$,*

$$P\left(\Delta_{emp}(x,C_p) \geq (\gamma - \epsilon) \cdot \Delta_{emp}(x,C_q)\right) \geq 1 - 3\exp\left(-\left(\frac{\epsilon(\gamma-1)\alpha}{\sqrt{8}\gamma(\gamma^2+1)}\right)^2 \cdot m\right)$$

The proof follows by a standard application of the Hoeffding bound and is deferred to the appendix.

Corollary 2.4 *Let $S = \{Z_1, \ldots, Z_m\}$ be an i.i.d. sample with distribution P. Then, for every (α,γ)-clustering $\mathcal{C}$, $\Pr(d(\mathcal{C}, \mathcal{C}^{\gamma-\delta}) > t) \leq \frac{3}{t\alpha} \cdot \exp\left(-\left(\frac{(\gamma-1)\delta\alpha}{\sqrt{8}\gamma(\gamma^2+1)}\right)^2 \cdot m\right)$.*

Proof. Denote $\mathcal{C} = \{C_1, \ldots, C_k\}$. By lemma 2.3, with $\epsilon = \delta$, we have

$$\begin{aligned}
E[d(C, C^{\gamma-\delta})] &= E[P(\cup_{i=1}^k C_i \oplus C_i^{\gamma-\delta})] \\
&= \sum_{i=1}^k \int_{C_i} \Pr(x \notin C_i^{\gamma-\delta}) dP(x) \\
&= \sum_{i=1}^k \sum_{j \neq i} \int_{C_i} \Pr(x \in C_j^{\gamma-\delta}) dP(x) \\
&\leq \sum_{i=1}^k (k-1) \cdot P(C_i) \cdot 3 \cdot \exp\left(- \left(\frac{(\gamma-1)\delta\alpha}{\sqrt{8}\gamma(\gamma^2+1)} \right)^2 \cdot m \right) \\
&= (k-1) \cdot 3 \cdot \exp\left(- \left(\frac{(\gamma-1)\delta\alpha}{\sqrt{8}\gamma(\gamma^2+1)} \right)^2 \cdot m \right)
\end{aligned}$$

Thus, the lemma follows from Markov's inequality and the fact that $k - 1 \leq k \leq \frac{1}{\alpha}$ □

We next turn to investigate the collection of all (α, γ)-clusterings of the given space. We observe first that every two distinct (α, γ)-clusterings must differ substantially.

Lemma 2.5 *If $\mathcal{C}, \mathcal{C}'$ are two (α, γ)-clusterings with $d(\mathcal{C}, \mathcal{C}') > 0$, then $d(\mathcal{C}, \mathcal{C}') \geq \frac{\alpha \cdot (\gamma-1)^2}{2\gamma^2 - \gamma + 1}$.*

Proof. Denote $\mathcal{C} = \{C_1, \ldots, C_k\}$, $\mathcal{C}' = \{C'_1, \ldots, C'_{k'}\}$ and $\epsilon = d(\mathcal{C}, \mathcal{C}')$. By adding empty clusters if needed, we can assume that $k = k'$. By reordering the clusters, if necessary, we can assume that $P(\cup_{i=1}^k C_i \oplus C'_i) = \epsilon$ and $P(C'_1 \oplus C_1) > 0$. Again by selecting the ordering we can assume the existence of some point x that is in $C'_1 \setminus C_1$ and in $C_2 \setminus C'_2$.

$$\begin{aligned}
\Delta(x, C'_1) &= \frac{1}{P(C'_1)} \cdot \int_{C'_1} d(x, y) dP(y) \\
&\geq \frac{1}{P(C'_1)} \cdot \int_{C_1} d(x, y) dP(y) - \frac{1}{P(C'_1)} \cdot \int_{C_1 \setminus C'_1} d(x, y) dP(y) \\
&\geq \frac{P(C_1)}{P(C'_1)} \cdot \Delta(x, C_1) - \frac{P(C_1 \setminus C'_1)}{\alpha} \cdot \max_{y \in C_1 \setminus C'_1} d(x, y) \qquad (1) \\
&\geq \left(1 - \frac{\epsilon}{\alpha}\right) \cdot \Delta(x, C_1) - \frac{\epsilon}{\alpha} \cdot \frac{\gamma^2+1}{\gamma(\gamma-1)} \Delta(x, C_1) \\
&\geq \left(1 - \frac{\epsilon}{\alpha} \cdot \frac{2\gamma^2 - \gamma + 1}{\gamma(\gamma-1)}\right) \cdot \gamma \cdot \Delta(x, C_2)
\end{aligned}$$

For the second inequality note that $\frac{P(C'_1)}{P(C_1)} \geq \frac{P(C_1) - P(C_1 \setminus C'_1)}{P(C_1)} \geq 1 - \frac{\epsilon}{\alpha}$. The third inequality follows from lemma 2.2.

As we just saw $\frac{\Delta(x,C_1')}{\Delta(x,C_2)} \geq \left(1 - \frac{\epsilon}{\alpha} \cdot \frac{2\gamma^2-\gamma+1}{\gamma(\gamma-1)}\right) \cdot \gamma$. The same argument yields as well $\frac{\Delta(x,C_2)}{\Delta(x,C_1')} \geq \left(1 - \frac{\epsilon}{\alpha} \cdot \frac{2\gamma^2-\gamma+1}{\gamma(\gamma-1)}\right) \cdot \gamma$. Consequently $1 \geq \left(1 - \frac{\epsilon}{\alpha} \cdot \frac{2\gamma^2-\gamma+1}{\gamma(\gamma-1)}\right) \cdot \gamma$ which proves the lemma. □

As we observe next, for every $\alpha > 0$ and $\gamma > 1$ the number of (α, γ)-clusterings that any space can have does not exceed $f(\alpha, \gamma)$, where f depends only on α and γ but *not* on the space. We find this somewhat surprising, although the proof is fairly easy.

Corollary 2.6 *There is a function $f = f(\alpha, \gamma)$ defined for $\alpha > 0$ and $\gamma > 1$ with the following property. The number of (α, γ)-clusterings of* any *metric probability space X is at most $f(\alpha, \gamma)$. This works in particular with $f(\alpha, \gamma) = 2 \cdot \left(\frac{12(2\gamma^2-\gamma+1)}{\alpha^2(\gamma-1)^2}\right)^{\left(\frac{\sqrt{8}\gamma(\gamma^2+1)}{(\gamma-1)^2\alpha}\right)^2 \cdot \ln(\frac{1}{\alpha})}$*

Proof. Consider the following experiment. We take an i.i.d. sample $Z_1, \ldots, Z_m$ of points from the distribution P with

$$m > \left(\frac{\sqrt{8}\gamma(\gamma^2+1)}{(\gamma-1)^2\alpha}\right)^2 \cdot \ln\left(\frac{12(2\gamma^2-\gamma+1)}{\alpha^2(\gamma-1)^2}\right).$$

and partition them randomly into $k \leq (\frac{1}{\alpha})$ parts $T_1, \ldots, T_k$. This induces a partition $\mathcal{C}^* = \{C_1, \ldots, C_k\}$ of the space X defined by

$$C_i = \{x \in X : \forall j \neq i,\ \Delta_U(x, T_i) < \Delta_U(x, T_j)\}$$

For every (α, γ)-clustering $\mathcal{C}$ of X we consider the event $A_{\mathcal{C}}$ that the induced partition of X satisfies $d(\mathcal{C}, \mathcal{C}^*) < \alpha \cdot \frac{(\gamma-1)^2}{2\cdot(2\gamma^2-\gamma+1)}$. Let us consider the events $A_{\mathcal{C}}$ over distinct (α, γ)-clusterings of the space. By Lemma 2.5, these events $A_{\mathcal{C}}$ are disjoint. Now consider the event B that the T_i's are consistent with $\mathcal{C}$. There are at most $(\frac{1}{\alpha})^m$ ways to partition the sampled points into $\frac{1}{\alpha}$ parts or less, so that $\Pr(B) \geq \alpha^m$. By the choice of m and by Corollary 2.4 $\Pr(A_{\mathcal{C}}|B) \geq \frac{1}{2}$. Thus, $\Pr(A_{\mathcal{C}}) \geq \Pr(B) \cdot \Pr(A_{\mathcal{C}}|B) \geq \frac{1}{2}\alpha^m$. Consequently, X has at most $f(\alpha, \gamma) = 2(\frac{1}{\alpha})^m$ distinct (α, γ)-clusterings, as claimed. □

Note 2.7 *Fix $\alpha > 0$. The number of (α, γ)-clusterings might be quite large when γ is close to 1. For example, let X be an n-point space, with uniform metric and uniform probability measure. Every partition in which each part has cardinality $\geq \alpha \cdot n$ is an $(\alpha, \frac{n}{n-1})$-clustering*[4].

[4]Note that this example is not valid if we define $\Delta(x, A) = E[d(x, y)|y \in A \setminus \{x\}]$. To overcome this point, we can replace every point $x \in X$ by many copies, where two copies of x are distance ϵ and a copy of x and a copy of $y \neq x$ are at distance $d(x, y)$.

Algorithmic Aspects

Fix $\alpha > 0, \gamma > 1$. We shall now show that an (α, γ)-clustering can be well approximated efficiently. By lemma 2.4, (α, γ)-clustering can be approximated by a small sample, where the approximation is with respect to the symmetric difference metric. A major flaw of this approximation scheme is that we have no verification method to accompany it. We do not know how to check whether a given partition is close to an (α, γ)-clustering w.r.t. the symmetric difference metric. To this end, we introduce another notion of approximation. A family of subsets of X, $\mathcal{C} = \{C_1, \ldots, C_k\}$, is an $(\epsilon, \alpha, \gamma)$-**clustering** if

- For every $i \in [k]$, $P(C_i) \geq \alpha$
- There is a set $N \subset X$ with $P(N) \leq \epsilon$ such that every $x \in X \setminus N$, belongs to exactly one C_i and for every $j \neq i$, $\Delta(x, C_j) \geq \gamma \cdot \Delta(x, C_i)$.

We consider next a partition that is attained by the method of Corollary 2.4. We show that if it is ϵ-close to an (α, γ)-clustering w.r.t. symmetric differences, then it is necessarily an $(\alpha - \epsilon, \gamma - O(\epsilon), \epsilon)$-clustering.

We associate with every collection $\mathcal{A} = \{A_1, \ldots, A_k\}$ of finite subsets[5] of X the following collection of subsets $\mathcal{C}^\gamma(\mathcal{A}) = \{C_1^\gamma(\mathcal{A}), \ldots, C_k^\gamma(\mathcal{A})\}$:

$$C_i^\gamma(\mathcal{A}) = \{x \in X : \forall j \neq i,\ \gamma \cdot \Delta_U(x, A_i) < \Delta_U(x, A_j)\} \tag{2}$$

where, as above, $\Delta_U(x, A) := \frac{1}{|A|} \sum_{z \in A} d(x, z)$.

Proposition 2.8 *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be an (α, γ)-clustering. Let $\mathcal{A} = \{A_1, \ldots, A_k\}$ where $\forall i,\ A_i \subset C_i$ and $d(\mathcal{C}^\gamma(\mathcal{A}), \mathcal{C}) < \epsilon$. Then $\mathcal{C}^\gamma(\mathcal{A})$ is an $(\alpha - \epsilon, \gamma - O(\epsilon), \epsilon)$-clustering. The unspecified coefficients in the O-term depend on α and γ.*

The main idea of the proof is rather simple: The assumption $d(\mathcal{C}^\gamma(\mathcal{A}), \mathcal{C}) < \epsilon$ implies that for all i the set $C_i \oplus C_i^\gamma(\mathcal{A})$ is small. This suggests that $\Delta(x, C_i) \approx \Delta(x, C_i^\gamma(\mathcal{A}))$ for most points $x \in X$. The only difficulty in realizing this idea is that points in $C_i \oplus C_i^\gamma(\mathcal{A})$ might have a large effect on either $\Delta(x, C_i)$ or $\Delta(x, C_i^\gamma(\mathcal{A}))$. But the assumption that $A_i \subset C_i$ gives us control over the distances between x to these points. The full proof can be found in the appendix.

To recap, the above discussion suggests a randomized algorithm that for a given $\epsilon > 0$ runs in time poly$(\frac{1}{\epsilon})$ and finds w.h.p. an $(\alpha - \epsilon, \gamma - O(\epsilon), \epsilon)$-clustering of X provided that X has an (α, γ)-clustering $\mathcal{C}$. We take $m = \Theta(\log(\frac{1}{\epsilon}))$ i.i.d. samples from X and go over all possible partitions of the sample points into at most $\frac{1}{\alpha}$ sets. There are only $\left(\frac{1}{\epsilon}\right)^{O(\log(\frac{1}{\alpha}))}$ such partitions. We next check whether the clustering of X that is induced as in Equation (2) is an $(\alpha - \epsilon, \gamma - O(\epsilon), \epsilon)$-clustering (this can be easily done by standard statistical estimates).

[5] In fact, we will allow $A_1, \ldots, A_k$ to have multiple points. Formally, then, $A_1, \ldots, A_k$ are multisets.

To see that the algorithm accomplishes what it should, note that the failure probability in corollary 2.4 with $\delta = \gamma - 1$ can be $\leq \frac{1}{2}$ for $m = \Theta(\log(\frac{1}{\epsilon}))$. Thus, w.p. $> \frac{1}{2}$ one of the considered partitions induces a partition of X which is ϵ-close in the symmetric difference sense to $\mathcal{C}$. By Proposition 2.8, this partition is an $(\alpha - \epsilon, \gamma - O(\epsilon), \epsilon)$-clustering.

This also proves Theorem 1.1: If our input is a finite metric space X, we can apply the above algorithm with $\epsilon = \frac{1}{|X|+1}$ and examples that are being sampled from X uniformly at random. As explained, w.h.p., the algorithm will consider every partition which is ϵ-close in the symmetric difference sense to any of X's (α, γ)-clusterings. However, since $\epsilon = \frac{1}{|X|+1}$, two ϵ-close partitions must be identical. This proves Theorem 1.1.

Note that by corollary 2.6, *all* the (α, γ)-clusterings can be approximated. A similar algorithm can efficiently find an approximate $(\alpha, \gamma, \epsilon)$-clustering, provided that one exists[6]. Also, similar techniques yield an algorithm to approximate an individual (α, γ)-cluster.

3 Clustering into Many Clusters

To simplify matters we consider only finite metric spaces endowed with a uniform probability distribution[7].

Lemma 3.1 *Let X be a metric space and let $\epsilon > 0$.*

1. *Let $C_1, C_2 \subseteq X$ be two $(3+\epsilon)$-clusters. Then $C_1 \cap C_2 = \emptyset$, $C_1 \subset C_2$ or $C_2 \subset C_1$.*
2. *Every $(3+\epsilon)$-cluster is a ball around one of its points.*
3. *The claim is sharp and the above claims need not hold for $\epsilon = 0$.*

Proof. We prove the first claim by contradiction and assume that $P(C_1 \setminus C_2), P(C_2 \setminus C_1), P(C_1 \cap C_2)$ are positive. Let $x \in C_1 \cap C_2$, $y \in C_1 \oplus C_2$ be such that $d(x,y)$ is as small as possible. Say that $y \in C_2$. Clearly, $\Delta(x, C_1 \setminus C_2) \geq d(x,y)$.

[6] The main difference is that here we do not consider partitions of the whole sample set. Rather, we seek first those sample points that belong to the exceptional set, and only partitions of the remaining sample points are considered.

[7] As in the previous section, it's a fairly easy matter to accommodate general metric spaces and arbitrary probability distributions.

We first deal with the case $P(C_1 \setminus C_2) \geq P(C_1 \cap C_2)$, and arrive at a contradiction as follows:

$$\begin{aligned}
\Delta(x, C_1) &= \frac{P(C_1 \setminus C_2)}{P(C_1)} \Delta(x, C_1 \setminus C_2) + \frac{P(C_1 \cap C_2)}{P(C_1)} \Delta(x, C_1 \cap C_2) \\
&\geq \frac{1}{2} \Delta(x, C_1 \setminus C_2) \\
&\geq \frac{1}{2} d(x, y) \\
&\geq \frac{1}{2} [\Delta(y, C_1) - \Delta(x, C_1)] \\
&\geq \frac{3 + \epsilon - 1}{2} \Delta(x, C_1)
\end{aligned}$$

When $P(C_1 \setminus C_2) \leq P(C_1 \cap C_2)$, a contradiction is reached as follows. By the choice of x, y, for every $z \in C_1 \setminus C_2$, there holds $\Delta(z, C_1 \cap C_2) \geq d(x, y)$. Therefore,

$$\begin{aligned}
\Delta(z, C_1) &= \frac{P(C_1 \setminus C_2)}{P(C_1)} \Delta(z, C_1 \setminus C_2) + \frac{P(C_1 \cap C_2)}{P(C_1)} \Delta(z, C_1 \cap C_2) \\
&\geq \frac{1}{2} \Delta(z, C_1 \cap C_2) \\
&\geq \frac{1}{2} d(x, y) \\
&\geq \frac{1}{2} [\Delta(y, C_1) - \Delta(x, C_1)] \\
&\geq \frac{1}{2} \cdot (1 - \frac{1}{3 + \epsilon}) \cdot \Delta(y, C_1) \\
&\geq \frac{1}{2} \cdot (1 - \frac{1}{3 + \epsilon}) \cdot (3 + \epsilon) \cdot \Delta(z, C_1)
\end{aligned}$$

To prove the second part, let C be a $(3 + \epsilon)$-cluster of diameter r, and let $x, y \in C$ satisfy $d(x, y) = r$. Since $d(x, y) \leq \Delta(x, C) + \Delta(y, C)$, we may assume w.l.o.g. that $\Delta(x, C) \geq \frac{d(x,y)}{2}$. We show now that $C = B(x, r)$ and C is a ball, as claimed. Indeed $d(x, z) \leq r$ for every $z \in C$, and if $z \notin C$, then $d(x, z) \geq \Delta(z, C) - \Delta(x, C) \geq (3 + \epsilon - 1)\Delta(x, C) > d(x, y) = r$. The conclusion follows.

To show that the result is sharp, consider the graph G that is a four-vertex cycle and its graph metric. It is not hard to check that every two consecutive vertices in G constitute a 3-cluster which is not a ball. Moreover a pair of intersecting edges in G yield an example for which the first part of the lemma fails to hold. □

An (α, γ)-cluster in a space X is called *minimal* if it contains no (α, γ)-cluster other than itself. Such clusters are of interest, since they can be viewed as "atoms" in clustering X.

Corollary 3.2 *For every $\alpha, \epsilon > 0$ and every space X there is at most one partition of X into minimal $(\alpha, 3 + \epsilon)$-clusters.*

To see this, consider two $(\alpha, 3+\epsilon)$-clusters C and C' that belong to two different such partitions and have a nonempty intersection. By Lemma 3.1, they must be comparable. By the minimality assumption, $C = C'$ which proves the claim.

Note 3.3 *We note that the previous Corollary may fail badly without the minimality assumption. Let $X = \{x_1, \ldots, x_n\} \dot\cup \{y_1, \ldots, y_n\}$, where $d(x_i, y_i) = 1$ for all i and all other distance equal γ. It is not hard to see that the following are (α, γ)-clusters in X where $\alpha = \frac{1}{2n}$: A singleton and a pair $\{x_i, y_i\}$. There are $2^n = 2^{\frac{1}{2\alpha}}$ ways to partition X into such clusters.*

Algorithmic Aspects

We next discuss several algorithmic aspects of clustering into arbitrarily many clusters. Our input consists of a finite metric space X and the parameter $\alpha > 0$. Lemma 3.1 suggests an algorithm for finding $(\alpha, 3+\epsilon)$-clusters and for partitioning the space into $(\alpha, 3+\epsilon)$-clusters. The runtime of this algorithm is polynomial in $|X|$, and independent of α. The second part of the lemma suggests how to find all the $(\alpha, 3+\epsilon)$-clusters. As the first part of the lemma shows, the inclusion relation among the $(\alpha, 3+\epsilon)$-clusters has a tree structure. Thus, we can use dynamic programming to find a partition of the space into $(\alpha, 3+\epsilon)$-clusters, provided that such a partition exists. This proves the positive part of Theorem 1.2.

To match the above positive result, we show

Theorem 3.4 *The following problems are NP-Hard.*

1. *$(\alpha, 2.5)$-CLUSTERING: Given an n-point metric space X and $\alpha > 0$, decide whether X has a $(\alpha, 2.5)$-clustering.*

2. *PARTITION-INTO-$(\alpha, 2.5)$-CLUSTERS: Given an n-point metric space X and $\alpha > 0$, decide whether X has a partition into $(\alpha, 2.5)$-clusters.*

The proof of this Theorem, which also proves the negative part of Theorem 1.2, is deferred to the appendix.

4 Conclusion

4.1 Relation to other work

As we explain below, our work is inspired by the classical VC/PAC theory. In addition we refer to several recent papers that contribute to the development of a theory of clustering.

VC/PAC theory

The VC/PAC setting offers the following formal description of the classification problem. We are dealing with a space $\mathcal{X}$ of *instances*. The problem is to recover an unknown member h^* in a known class $\mathcal{H}$ of hypotheses. Here $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, where $\mathcal{Y}$ is a finite set of *labels*. We seek to recover the unknown h^* by observing a sample $S = \{(x_i, h^*(x_i)\}_{i=1}^m \subset \mathcal{X} \times \mathcal{Y}$. These samples come from some fixed but unknown distribution over $\mathcal{X}$.

Our description of the clustering problem is similar. We consider a space X of instances and a class $\mathcal{G}$ of good clusterings $(P, \mathcal{C})$ of X, where P is probability measure over X and $\mathcal{C}$ is a partition of X. We are given a sample $\{X_1, \ldots, X_m\} \subset X$ that comes from some unknown P, where $(P, \mathcal{C}) \in \mathcal{G}$ for some partition $\mathcal{C}$, and our purpose is to recover $\mathcal{C}$. Specifically, here X is a metric space, $\mathcal{G}$ is the class of probability measures P that admit a partition which is a (α, γ)-clustering and the corresponding partition is the associated (α, γ)-clustering.

Both theories seek conditions on $\mathcal{G}$ or $\mathcal{H}$ under which there are no information theoretic or computational obstacles that keep us from performing the above mentioned tasks.

Alternative Notions of Good Clustering

Our approach is somewhat close in spirit to [4], see also [6]. These papers assume that the space under consideration has a clustering with some structural properties, and show how to find it efficiently. In particular, a key notion in these papers is the **γ-average attraction property**, which is conceptually similar to our notion of γ-clustering. Given a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ of a space X it is possible to compare between clusters either additively or through multiplication. In [4] the requirement is that $\Delta(x, C_i) + \gamma \leq \Delta(x, C_j)$ for every $x \in C_i$ and $j \neq i$, whereas our condition is $\Delta(x, C_i) \cdot \gamma \leq \Delta(x, C_j)$. A clear advantage of our notion is its scale invariance. On the other hand, their algorithms work even if X is not a metric space and is only endowed with an arbitrary dissimilarity function.

We mention two more papers that share a similar spirit. Consider a data set that resides in the unit ball of a Hilbert Space. It is shown in [8] how to efficiently find a large margin classifier for the data provided that one exists. In [1] several additional possible notions of good clustering are introduced and analyzed.

Stability

The notion of instance stability was introduced in [5] (See also [3]). An instance for an optimization problem in called *stable* if the optimal solution does not change (or changes only slightly) upon a small perturbation of the input. The point is made that instances of clustering problems are of practical interest only if they are stable. The notion of an (α, γ)-clustering has a similar stability property. Namely, if we slightly perturb a metric, an (α, γ)-clustering is still (α', γ')-clustering for $\alpha' \approx, \alpha,\ \gamma' \approx \gamma$.

Thus, a good clustering remains a good clustering under a slight perturbation of the input

In fact, the present paper is an outgrowth of our work on stable instances for MAXCUT, which we view as a clustering problem. We recall that the input to the MAXCUT problem is an $n \times n$ nonnegative symmetric matrix W. We seek an $S \subseteq [n]$ which maximizes $\sum_{i \in S, j \notin S} w_{ij}$. Even METRIC-MAXCUT problem (i.e., when w_{ij} form a metric) is NP-Hard []. We say that W' is a γ-perturbation of W some $\gamma > 1$ if $\forall i, j,\ \gamma^{-\frac{1}{2}} w_{ij} \leq w'_{i,j} \leq \gamma^{\frac{1}{2}} w_{ij}$. The instance W of MAXCUT is called **γ-stable** if the optimal solution S for W coincides with the optimal solution for every γ-perturbation W' of W. The methods presented in this paper can be used to give, for every $\epsilon > 0$, an efficient algorithm that correctly solves all $(1 + \epsilon)$-stable instances of METRIC-MAXCUT.

These developments will be elaborated in a future publication.

4.2 Future Work and Open Questions

In view of this article and papers such as [1, 8, 4] it is clear that there is still much interest in new notions of a good clustering and the relevant algorithms. Still, on the subjects discussed here several natural questions remain open.

1. We believe that it should be possible to improve the dependence on α and γ of the run time of the algortihm in Theorem 1.1.

2. We gave an efficient method for partitioning a space into 3-clusters, and showed (theorem 3.4) that it is NP-Hard to find a partition into 2.5-clusters. Can this gap be closed?

3. As Lemma 3.1 shows, $(3 + \epsilon)$-clusters are just balls. It is not hard to see that Lemma 2.3 implies that given an (α, γ)-clustering of an n-point metric space, it is possible to find $O_\gamma(\log n)$ *representative* points in every cluster so that the clustering is nothing but the Voronoi diagram of the (bunched) representative sets. Presumably, there is still some interesting structural theory of (α, γ)-clustering waiting to be discovered here. Specifically, can the above $O_\gamma(\log n)$ be replaced by $O_\gamma(1)$? A positive answer would give a deterministic version of our algorithm from section 2, with no dependency of α, but only on the maximal number of clusters.

4. Consider the following statement "Every n-point metric space X has a partition $X = A \dot{\cup} B$ such that for every $x \in A, y \in B$, it holds that $\gamma(n) \cdot \Delta(x, A) \leq \Delta(x, B)$ and $\gamma(n) \cdot \Delta(y, B) \leq \Delta(y, A)$". How large can $\gamma(n)$ be for this statement to be true?

References

[1] M. Ackerman and S. Ben-David. Which data sets are "clusterable"? a theoretical study of clusterability. *NIPS*, 2008.

[2] Sanjeev Arora, Rong Ge, Sushant Sachdeva, and Grant Schoenebeck. Finding overlapping communities in social networks: Toward a rigorous approach. Technical report, 2012. http://arxiv.org/abs/1112.1831.

[3] P. Awasthi, A. Blum, and O. Sheffet. Center-based clustering under perturbation stability. In *Information Processing Letters*, volume 112, pages 49–54, 2011.

[4] M.F. Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *STOC*, pages 671–680, 2008.

[5] Y. Bilu and N. Linial. Are stable instances easy? In *ICS*, pages 332 – 341, 2010.

[6] Avrim Blum. Thoughts on clustering. In *NIPS Workshop on Clustering Theory*, 2009.

[7] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco, 1979.

[8] Z. S. Karnin, E. Liberty, S. Lovett, R. Schwartz, and O. Weinshtein. On the furthest hyperplane problem and maximal margin clustering. arXiv.org:1107.1358v1, 2011.

[9] J. Kleinberg. An impossibility theorem for clustering. *NIPS*, pages 446–453, 2002.

A Proofs omitted from the text

Proof. (of Lemma 2.3) For $A \subset X$, denote $I_A = \frac{1}{m}|\{j : Z_j \in A\}|$, $J_A = \sum_{j:Z_j \in A} d(x, Z_j)$. For every $j \in [m]$ define

$$Y_j = \begin{cases} \frac{1}{P(C_p)} \cdot d(x, Z_j) & Z_j \in C_p \\ -\frac{\gamma - \frac{\epsilon}{2}}{P(C_q)} \cdot d(x, Z_j) & Z_j \in C_q \\ 0 & otherwise \end{cases}$$

We have $EY_j = \Delta(x, C_p) - (\gamma - \frac{\epsilon}{2}) \cdot \Delta(x, C_q) \geq \frac{\epsilon}{2\gamma} \cdot \Delta(x, C_p)$. Moreover, by lemma 2.2, $|Y_j| \leq \frac{(\gamma - \frac{\epsilon}{2})}{\alpha} \cdot \frac{\gamma^2+1}{\gamma(\gamma-1)} \cdot \Delta(x, C_p) \leq \frac{\gamma^2+1}{\alpha(\gamma-1)} \cdot \Delta(x, C_p)$. Thus, by Hoeffding's bound,

$$P\left(\frac{J_{C_p}}{P(C_p)} \leq \left(\gamma - \frac{\epsilon}{2}\right) \cdot \frac{J_{C_q}}{P(C_q)}\right) = P\left(\sum_{j=1}^{m} Y_j \leq 0\right) \leq \exp\left(-\left(\frac{\epsilon(\gamma-1)\alpha}{\sqrt{8}\gamma(\gamma^2+1)}\right)^2 \cdot m\right)$$

Again by Hoeffding's bound, we have

$$P\left(\frac{I_{C_q}}{P(C_q)} \leq 1 - \frac{\epsilon}{4\gamma}\right) \leq \exp\left(-\left(\frac{\epsilon\alpha}{\sqrt{8}\gamma}\right)^2 \cdot m\right)$$

$$P\left(\frac{I_{C_p}}{P(C_p)} \geq 1 + \frac{\epsilon}{4\gamma}\right) \leq \exp\left(-\left(\frac{\epsilon\alpha}{\sqrt{8}\gamma}\right)^2 \cdot m\right)$$

Combining the inequalities, we conclude that, with probability $\geq 1 - 3\exp\left(-\left(\frac{\epsilon(\gamma-1)\alpha}{\sqrt{8}\gamma(\gamma^2+1)}\right)^2 \cdot m\right)$,

$$\begin{aligned}\frac{\frac{J_{C_p}}{I_{C_p}}}{(\gamma-\epsilon)\frac{J_{C_q}}{I_{C_q}}} &= \frac{\frac{J_{C_p}}{P(C_p)}}{(\gamma-\frac{\epsilon}{2})\frac{J_{C_q}}{P(C_q)}} \cdot \frac{\frac{P(C_p)}{I_{C_p}}}{\frac{P(C_q)}{I_{C_q}}} \cdot \frac{\gamma-\frac{\epsilon}{2}}{\gamma-\epsilon} \\ &\geq \frac{1-\frac{\epsilon}{4\gamma}}{1+\frac{\epsilon}{4\gamma}} \cdot \frac{\gamma-\frac{\epsilon}{2}}{\gamma-\epsilon} \geq 1\end{aligned}$$

□

Proof (of Proposition 2.8) It is very suggestive how to select the exceptional set in the $(\alpha-\epsilon, \gamma-O(\epsilon), \epsilon)$-clustering that we seek. Namely, let $N = \cup_i (C_i \setminus C_i^\gamma(\mathcal{A}))$. As needed, $P(N) < \epsilon$, since $d(\mathcal{C}^\gamma(\mathcal{A}), \mathcal{C}) < \epsilon$. To prove our claim, note that $\forall i,\ P(C_i^\gamma) \geq \alpha - \epsilon$ since $d(\mathcal{C}, \mathcal{C}_*^\gamma(\mathcal{A})) < \epsilon$. Consider some $x \in X \setminus N$ and the unique index i for which $x \in C_i^\gamma(\mathcal{A}) \cap C_i$. If $j \neq i$, we need to show that

$$\Delta(x, C_j^\gamma(\mathcal{A})) \geq (\gamma - O(\epsilon))\Delta(x, C_i^\gamma(\mathcal{A}))$$

As in the proof of lemma 2.5, we have

$$\begin{aligned}\Delta(x, C_j^\gamma(\mathcal{A})) &\geq \left(1 - \frac{\epsilon}{\alpha}\right)\Delta(x, C_j) - \frac{\epsilon}{\alpha}\max_{y \in C_j \setminus C_j^\gamma(\mathcal{A})} d(x,y) \\ &\geq \left(1 - \frac{\epsilon}{\alpha} \cdot \frac{2\gamma^2 - \gamma + 1}{\gamma(\gamma-1)}\right) \cdot \Delta(x, C_j) \qquad (3) \\ &=: (1 - a_1 \cdot \epsilon) \cdot \Delta(x, C_j)\end{aligned}$$

Similarly, again as in the proof of lemma 2.5, we have

$$\Delta(x, C_i) \geq \left(1 - \frac{\epsilon}{\alpha}\right)\Delta(x, C_i^\gamma(\mathcal{A})) - \frac{\epsilon}{\alpha}\max_{y \in C_i^\gamma(\mathcal{A}) \setminus C_i} d(x,y) \qquad (4)$$

Now, for $y \in C_i^\gamma(\mathcal{A})$, we have

$$\begin{aligned}d(x,y) &\leq \Delta_U(x, A_i) + \Delta_U(y, A_i) \\ &\leq \frac{1}{\gamma}\Delta_U(x, A_j) + \frac{1}{\gamma}\Delta_U(y, A_j) \\ &\leq \frac{2}{\gamma}\Delta_U(x, A_j) + \frac{1}{\gamma}d(x,y)\end{aligned}$$

Now, since $A_j \subset C_j$, by lemma 2.2, $\Delta_U(x, A_j) \leq \frac{\gamma^2+1}{\gamma(\gamma-1)}\Delta(x, C_j)$ and we have,

$$d(x, y) \leq \frac{\gamma}{\gamma - 1} \cdot \frac{\gamma^2 + 1}{\gamma(\gamma - 1)} \cdot \frac{2}{\gamma}\Delta(x, C_j)$$

So, by equation (4) we have,

$$\Delta(x, C_i) \geq (1 - a_2 \cdot \epsilon) \cdot \Delta(x, C_i^{\gamma}(\mathcal{A})) - a_3 \cdot \epsilon \cdot \Delta(x, C_j) \tag{5}$$

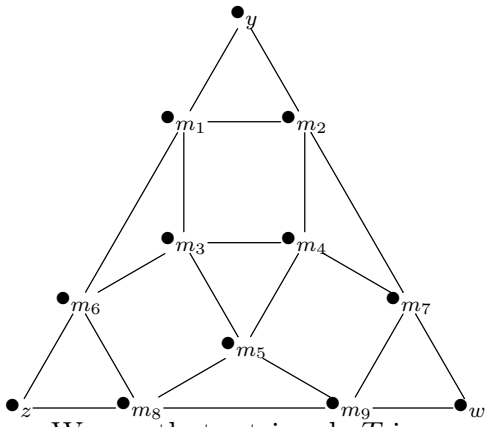
For some positive constants a_2, a_3 which depend only of γ and α. Now by equations (3) and (5) we conclude that

$$\begin{aligned}
\Delta(x, C_j^{\gamma}(\mathcal{A})) &\geq (1 - (a_1 + \gamma a_3) \cdot \epsilon) \cdot \Delta(x, C_j) + \gamma a_3 \cdot \epsilon \cdot \Delta(x, C_j) \\
&\geq (1 - (a_1 + \gamma a_3) \cdot \epsilon) \cdot \gamma \cdot \Delta(x, C_i) + \gamma a_3 \cdot \epsilon \cdot \Delta(x, C_j) \\
&\geq (1 - (a_1 + \gamma a_3) \cdot \epsilon)(1 - a_2 \cdot \epsilon)\gamma \cdot \Delta(x, C_i^{\gamma}(\mathcal{A})) \\
&= (\gamma - O(\epsilon)) \cdot \Delta(x, C_i^{\gamma}(\mathcal{A}))
\end{aligned}$$

Proof. (of Theorem 3.4) Both claims are proved by the same reduction from 3-DIMENSIONAL-MATCHING (e.g., [7] pp. 221). The input to this problem is a subset $M \subset Y \times Z \times W$, where Y, Z, W are three disjoint q-element sets. A *three dimensional matching* (=3DM) is a q-element subset $M' \subset M$ that covers all elements in $Y \dot{\cup} Z \dot{\cup} W$. The problem is to decide whether a 3DM exists.

We associate with this instance of the problem a graph on vertex set $Y \dot{\cup} Z \dot{\cup} W$, and edge set the union of all triangles $\{y, z, w\}$ over $(y, z, w) \in M$. It is not hard to see that $3DM$ remains NP-Hard under the restriction that this graph is connected.

Here is our reduction. Given an instance $M \subset Y \times Z \times W$ of $3DM$, we construct a graph $G^M = (V^M, E^M)$ as follows: Associated with every $m = (y, z, w) \in M$ is a gadget below. We consider the clustering problem on G^M with its natural graph metric.



We say that a triangle T in a graph is *isolated* if every vertex outside it has at most one neighbor in T. The above gadget is useful for the reduction since it's easy to verify that:

Claim 1 *The graph G^M can be partitioned into isolated triangles iff M has a $3DM$.*

Proof(sketch). If M has a $3DM$, we can construct a partition of V into isolated triangle by taking the triangles

$$\{y, m_1, m_2\}, \{z, m_6, m_8\}, \{w, m_7, m_9\}, \{m_3, m_4, m_5\} \tag{6}$$

for every m in the $3DM$ and the triangles

$$\{m_1, m_3, m_6\}, \{m_2, m_4, m_7\}, \{m_5, m_8, m_9\} \tag{7}$$

for m outside it. On the other hand, consider any partition of G^M into isolated triangles. Its restriction to every gadget must coincide with one of the above two choices, so that the corresponding $3DM$ is readily apparent □

Both NP-Hardness claims in Theorem 3.4 follow from the above discussion and the following claim

Claim 2 *Let $G = (V, E)$ be a connected graph in which all vertex degrees are ≥ 2. For every partition of the vertex set $V = \dot{\cup}_1^k C_i$, the following are equivalent*

1. *Each C_i induces an isolated triangle.*
2. *Each C_i is a $(\frac{3}{|V|}, 2.5)$-cluster.*
3. *The partition $C_1, \ldots, C_k$ is a $(\frac{3}{|V|}, 2.5)$-clustering.*

Proof The implication 1. $\Rightarrow$ 2. and 1. $\Rightarrow$ 3. are easily verified. We turn to prove 3. $\Rightarrow$ 1. Let $i \in [k]$. We need to show that each C_i is an isolated triangle. Clearly, $|C_i| \geq 3$ by definition of $(\frac{3}{|V|}, 2.5)$-clustering. But G is connected, so there are two neighbors xy with $x \in C_i, y \notin C_i$. By proposition 2.2 we have

$$1 = d(x, y) \geq (2.5 - 1)\Delta(x, C_i) \geq 1.5 \cdot \frac{|C_i| - 1}{|C_i|},$$

so that $|C_i| = 3$. Consider now $x, y \in C_i$ which are nonadjacent. Since $d(x) \geq 2$, it has a neighbor $z \notin C_i$. Using Proposition 2.2 we arrive at the following contradiction: $1 = d(x, z) \geq (2.5 - 1)\Delta(x, C_i) \geq 1.5 \cdot \frac{2+1}{3} = 1.5$. We already know that each C_i is a triangle, but why is it isolated? If $z \in C_j$, $j \neq i$ has at least two neighbors in C_i, then

$$2.5 \cdot \Delta(z, C_j) \leq \Delta(z, C_i) \leq \frac{4}{3} < 2.5 \cdot \frac{2}{3} = 2.5 \cdot \Delta(z, C_j).$$

The proof of 2. $\Rightarrow$ 1. is similar. Let C_i be cluster in the partition. Using the same argument as before, where the the fact that $\forall x \in C_i, y \notin C_i$, $d(x, y) \geq \Delta(y, C_i) - \Delta(y, C_i) \geq (2.5 - 1)\Delta(x, C_i)$ replace Proposition 2.2, we deduce that C_i induces a triangle. To show that C_i is isolated, suppose that there exists a vertex $z \notin C_i$ with ≥ 2 neighbors in C_i. Let $x \in C_i$ be an arbitrary vertex. To obtain a contradiction, we note that

$$2.5\Delta(x, C_i) \leq \Delta(z, C_i) \leq \frac{4}{3} < 2.5 \cdot \frac{2}{3} = 2.5\Delta(x, C_i)$$

□

Note A.1 *Theorem 3.4 is tight in the following sense: As the proof shows, the above problems are hard even for graph metrics. On the other hand, given a graph $G = (V, E)$, the following polynomial time algorithms find (i) A partition into $(\alpha, 2.5 + \epsilon)$-clusters, and (ii) A $(\alpha, 2.5 + \epsilon)$-clustering. (provided, of course, that one exists).*

1. *If $\alpha > \frac{1}{|V|}$ then, as in the proof of theorem 3.4, one shows that a partition into $(\alpha, 2.5 + \epsilon)$-clusters / $(\alpha, 2.5 + \epsilon)$-clustering is equivalent to a perfect matching, no edge of which is contained in a triangle. This can be done by first eliminating every edge that belongs to a triangle and then running an arbitrary matching algorithm.*

2. *If $\alpha < \frac{1}{|V|}$ then clearly there is no partition into $(\alpha, 2.5+\epsilon)$-clusters / $(\alpha, 2.5+\epsilon)$-clustering. If $\alpha = \frac{1}{|V|}$, the singletons constitute a partition of V into $(\alpha, 2.5 + \epsilon)$-clusters and a $(\alpha, 2.5 + \epsilon)$-clustering.*

Note A.2 *As in Note 2.7, by replacing each vertex with many points at distance ϵ from each other, the above reduction applies as well with the definition $\Delta(x, A) = E[d(x, y)|y \in A \setminus \{x\}]$.*